HIERARCHICAL DEM REFINEMENT USING SURFACE PARALLAX

Amit K Agrawal, Reuven Meth and Rama Chellappa
University of Maryland
College Park, MD - 20770
{aagrawal,meth,rama}@cfar.umd.edu

ABSTRACT

We present a multi-resolution approach to update and refine coarse 3D models of urban environments from a sequence of intensity images using surface parallax. A coarse and potentially incomplete depth map of the scene obtained from a Digital Elevation Map (DEM) is used as a reference surface which is refined and updated using this approach. We first estimate the camera motion using the reference depth map. Using the estimated camera motion, at each level in the multi-resolution framework, motion of 3D points on the reference surface is compensated, and the residual flow field, which is an epipolar field, is estimated and used to refine the depth map at that level. At a coarse resolution, the difference between the reference depth and the true depth will be small, leading to a small parallax field. The refined depth map from the coarser level is then propagated to the finer level and is used as a reference depth map at that level. Thus, significant deviations of an available model from a true model can be handled using this approach.

1 INTRODUCTION

There has been considerable interest recently in using autonomous mobile robots in surveillance. The ability to send mobile, sensor-equipped robots into environments that are potentially hazardous to humans is of vital importance in a number of scenarios (e.g. nuclear/biological/chemical contamination). There is a need for robust, real-time algorithms that exploit data collected by sensors mounted on the robots in order to improve the operators awareness of the scene. The operator's control station often has access to some meta-data, e.g. elevation data of the environment in which the robots are operating. In such a situation, it would be very useful to be able to integrate video from the robots with elevation data to provide the operator with a more accurate picture of the environment. Elevation data is often available in the form of a Digital Elevation Map (DEM), which gives the elevation of terrain over a geographical area. Thus the available DEM can be used to obtain the reference surface (depth map) of the scene. In general, these depth maps are coarse and may contain partial information about the area due to structural changes (e.g. construction, demolition of buildings). This coarse reference surface can be updated and refined using information from a sequence of 2D images of the scene. The enhanced scene can provide a remote operator a better understanding of the scene in which robot is operating. In addition, changes in urban environments such as addition of new buildings, demolition of old buildings or other structural changes, can be incorporated in the DEM without requiring additional dedicated DEM data collection.

2 THEORY

For any two views of a scene under perspective projection, if the motion of the 3D points on a surface is compensated, the resulting parallax field is an epipolar field. Referring to Figure 1, let C_1 and C_2 represent the camera center for two views and S be the reference surface which is aligned. Let Q be the 3D point on the reference surface, P be the true location of the 3D point and the projection of these points in reference image C_1 be q and p respectively. The residual parallax can be shown to be equal to (Kumar, 1994;Agrawal, 2004)

$$\delta u = q - p = \frac{T_z(Q_z - P_z)}{Q_z(P_z - T_z)}(p - e) \tag{1}$$

where e denotes the epipole and T_z denotes the translation in Z direction. If $T_z = 0$,

$$\delta u = q - p = \frac{-f(Q_z - P_z)}{Q_z P_z}(t) \tag{2}$$

where f is the focal length and $t = [T_x, Ty]^T$ denotes the 2×1 translation vector in x, y space. Without loss of generality, for the rest of the paper we assume $T_z > 0$.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
00 DEC 2004	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Hierarchical Dem Refinement Using Surface Parallax	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Maryland College Park, MD - 20770**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001736, Proceedings for the Army Science Conference (24th) Held on 29 November - 2 December 2004 in Orlando, Florida., The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**7**	

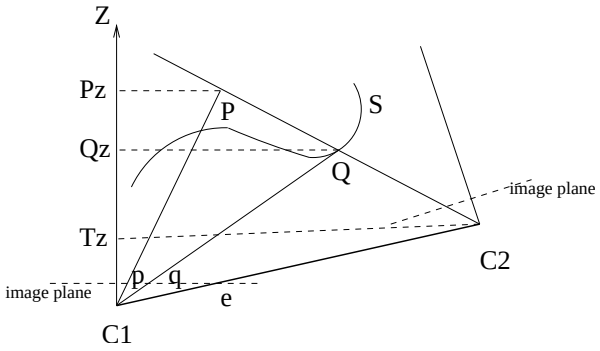


Figure 1: Parallax due to surface S

Since (1) has the unknown correspondence p on the right hand side, it is solved for parallax in terms of q as

$$\delta u = q - p = \frac{T_z(Q_z - P_z)}{P_z(Q_z - T_z)}(q - e) \tag{3}$$

Then $\beta = \frac{T_z(Q_z-P_z)}{P_z(Q_z-T_z)}\|(q-e)\|$ denotes the parallax magnitude and $\mathbf{v} = \frac{(q-e)}{\|(q-e)\|}$ denotes the parallax direction. The true depth P_z can be estimated using the parallax magnitude as

$$P_z = \frac{T_z Q_z}{\frac{\beta}{\|(q-e)\|}(Q_z - T_z) + T_z} \tag{4}$$

3 APPROACH

Our approach uses a hierarchical framework to align a non-planar surface (hereby referred to as the *reference surface*) in images and estimate the deviations from the reference surface by calculating the residual parallax field. The algorithm uses two frames from the image sequence, one of them being the reference frame for which the depth map is refined. For the rest of the paper, we refer to the reference image as *key* image and the second image as the *offset* image.

3.1 Estimating Camera Motion

We begin by first estimating the camera motion assuming that the camera calibration is known. We identify a small planar region in the 3D scene (orientation and distance in the camera coordinate system) using the reference depth map and its corresponding region in the key image. Since the optical flow of a planar surface is parametric (quadratic in image pixels), we fit a parametric optical flow (Bergen, 1992) to the region and obtain the parameters for that region. For a planar surface, the relationship among the optical flow parameters, the orientation and distance of the plane from the origin and motion parameters is well known (Trucco, 1998). Due to a coarse initial depth map, the motion parameters estimated using these equations may not be very accurate. However, they can be used as an initial estimate for refining the motion parameters as explained below.

Consider the equations relating the image motion of a rigid body with depth and camera motion (Horn, 1986)

$$\begin{aligned} u(x,y) =& \frac{-x_f + x}{Z'} + \frac{1}{f}xy\Omega_x \\ & - (f + \frac{1}{f}x^2)\Omega_y + y\Omega_z \end{aligned} \tag{5}$$

$$\begin{aligned} v(x,y) =& \frac{-y_f + y}{Z'} + (f + \frac{1}{f}y^2)\Omega_x \\ & - \frac{1}{f}xy\Omega_y - x\Omega_z \end{aligned} \tag{6}$$

where (x_f, y_f) denotes the FOE in image coordinates, $(\Omega_x, \Omega_y, \Omega_z)^T$ denotes the camera rotation velocities, Z' is the scaled depth, $Z' = \frac{Z_{ref}}{T_z}$, and $(u, v)^T$ denotes the 2-D velocities according to the reference depth Z_{ref}. We use the initial motion estimate $x_f, y_f, \Omega_x, \Omega_y, \Omega_z$ to estimate Z' and use the estimated Z' to refine the motion estimates. This is iterated until the motion estimates are stable or a specified number of iterations are reached. Finally, we obtain an estimate of T_z as $T_z = \frac{\langle Z_{ref} \rangle}{\langle Z' \rangle}$ where $\langle\rangle$ denotes the averaging operator over the planar surface.

Note that the FOE values are not affected by the estimated T_z because we are refining over the FOE values first and then estimating T_z using the refined scaled depths. In fact, since depth and T_z are coupled, we can use the scaled depth throughout our algorithm without explicitly computing T_z.

3.2 Hierarchical Framework

Let the superscript l denotes the resolution level, i.e. x^l denotes a variable at level l with $l = 1 \ldots L$ where

L is the coarsest level. Dividing each side of (3) by 2^l, we get

$$\begin{aligned}\frac{\delta u}{2^l} &= \frac{q}{2^l} - \frac{p}{2^l} = \frac{\frac{T_z}{2^l}(\frac{Q_z}{2^l} - \frac{P_z}{2^l})}{\frac{P_z}{2^l}(\frac{Q_z}{2^l} - \frac{T_z}{2^l})}(\frac{q}{2^l} - \frac{e}{2^l}) \\ &= \frac{\frac{T_z}{2^l}(Q_z^l - P_z^l)}{P_z^l(Q_z^l - \frac{T_z}{2^l})}(q^l - e^l)\end{aligned} \tag{7}$$

where $Q_z^l = \frac{Q_z}{2^l}$, $P_z^l = \frac{P_z}{2^l}$ denotes the assumed and true depth at level l and $q^l = \frac{q}{2^l}, e^l = \frac{e}{2^l}$ are the image pixel coordinates at level l. The above equation shows the relationship between the parallax and depths at level l. Thus we can see that at coarser levels, the parallax field is small. The hierarchical estimation algorithm proceeds as follows

1. Estimate the camera motion. Construct pyramids for the key and offset images and for the reference depth map.

2. Initialize $l = L$. Use the reference depth map and the motion estimates at coarsest level to estimate the parallax field (as described in the Appendix). Refine the depths using the parallax.

3. Propagate the depths to level $l - 1$. $l \rightarrow l - 1$.

4. Warp the offset image according to the propagated depth at the current level and use it to estimate the parallax field. Refine the depths using the estimated parallax.

5. Iterate steps 3 and 4 until $l = 1$.

4 EXPERIMENTS

We present results on both semi-synthetic and real world 3D models. In all experiments, images contain 640×480 pixels.

4.1 Semi-synthetic Models

For semi-synthetic models (with real textures), we rendered a 3D model of a city with buildings and objects in OpenGL. We simulated a sequence of images by moving a virtual camera in the scene. The depth maps were obtained from the OpenGL Z buffer. The depth maps are color coded (with brighter regions nearer to camera). Figures 2(a) and 2(b) show two frames from a synthetic image sequence respectively. Figure 2(c) shows the true depth map for the key frame and Figure 2(d) shows the reference depth map which was used as a surface for alignment. The background is kept at a depth of 1000 units. A portion of ground plane was used for camera motion estimation as explained in Section 3.1. The true and estimated camera motion parameters are as shown in Table 1 (with rotation angles in degrees).

Table 1: True and estimated motion parameters for semi-synthetic example

	T_x	T_y	T_z	W_x	W_y	W_z
True	-3.76	0.60	5.26	0.02	-1.29	1.47
Estimated	-3.68	0.57	4.93	0.01	-1.34	1.45

Table 2: Percentage depth error between the true depth map and the reference and estimated depth maps using different number of levels L

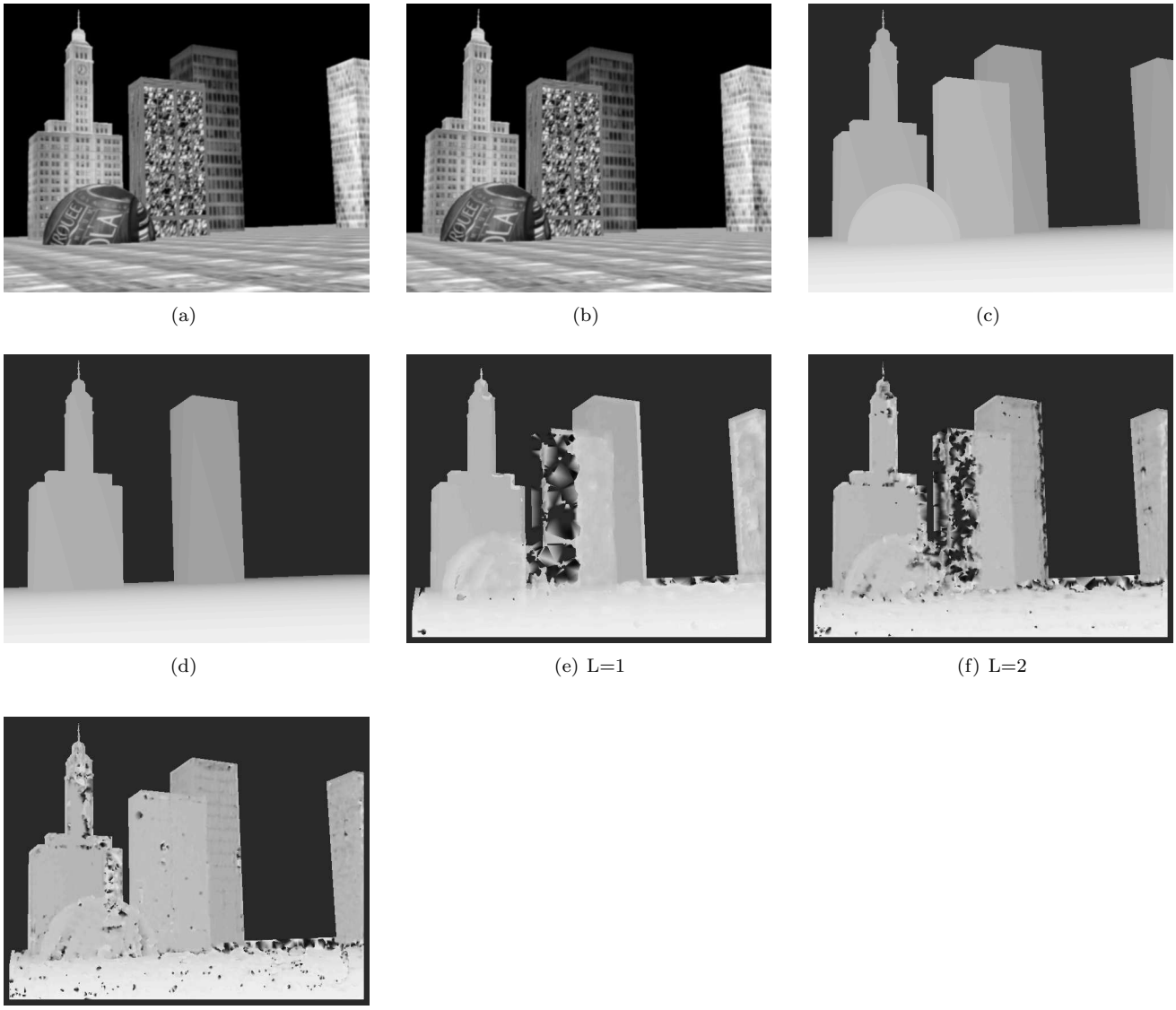
Depth Map	Percentage Depth Error
Reference	35.59
Estimated: L = 1	23.94
Estimated: L = 2	16.21
Estimated: L = 3	03.74

Figures 2(e), 2(f) and 2(g) show the estimated depth maps using different numbers of levels L in multiresolution framework. Notice that for $L = 1$, the depth of the portion of the building (in the center of the depth map image) which overlaps with the building at the back are estimated correctly, whereas for the portion which overlaps with the background, the depths are not estimated properly, because for pixels in that region the parallax magnitude due to high depth difference (from the background) is much higher. The maximum parallax magnitude at levels $1, 2$ and 3 are $10.37, 5.17$ and 2.57 pixels respectively. The estimated depth map using $L = 3$ is better than those obtained using $L = 1$ and 2.

We define the relative percentage depth error between the true depth map Z_{true} and some other depth map Z as $100 \times \frac{1}{N}\sum_1^N (\frac{Z_{true} - Z}{Z_{true}})^2$ where N denotes the total number of pixels in the image. Table 2 gives the percentage depth error between the true depth map and the initial reference and estimated depth maps using different numbers of levels L. Thus, the hierarchical approach was able to estimate the parallax for regions with high parallax magnitude. The results obtained using the hierarchical approach are much better both qualitatively and quantitatively.

4.2 Real World Models

A DEM model of downtown Baltimore (inner harbor area) was rendered in OpenGL and the reference depth



(a) (b) (c)

(d) (e) L=1 (f) L=2

(g) L=3

Figure 2: Semi-Synthetic example (a) Key frame (b) Offset frame (c) True depth map for key frame (d) Reference depth map used for alignment (e,f,g) Estimated depth maps using different levels L

map was obtained using the Z buffer as shown in Figure 3(c). The depth map is color coded (with brighter regions closer to camera). Video images were captured using a Sony camcoder placed on a cart (not mounted) moving across a street. Figures 3(a) and 3(b) show the key and offset frames from the video sequence respectively. Notice that the reference depth map is quite coarse. In order to show the effectiveness of the hierarchical framework, a portion of the reference depth map was modified to a very small depth value (shown in Figure 3(d)) so that the difference in depths for that portion of image is large leading to large parallax values. A portion of ground plane was used for camera motion estimation. Figures 3(e) and 3(f) show the estimated depths using the hierarchical algorithm for $L = 1$ and 3 respectively. Notice that for $L = 1$, the parallax field for the patch where the depths were modified to a low value is not estimated properly. As a result, the obtained depths are not correct (they are in fact much closer). For $L = 3$, we can see that a better estimate of depths is obtained. For example, the depth of the pole in the foreground is more accurately recovered.

CONCLUSIONS

A hierarchical framework for refining and updating a 3D model given a coarse depth map has been presented. The approach can be viewed as a fusion of available depth information (metadata) with the information from intensity images obtained from mobile robots. Results on both semi-synthetic 3D models and real models were presented. The estimated depth map is quite accurate for the semi-synthetic 3D model and appear plausible for the real model. The enhanced scene can provide a much better understanding of the scene in which robot is operating.

ACKNOWLEDGEMENTS

Prepared through collaborative participation in the Advanced Decision Architectures Consortium sponsored by the U. S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0009. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon. The authors would like to thank Phil David and Jeff DeHart of U. S. Army Research Laboratory for helpful discussions.

APPENIDX

Let $I(x, y, t)$ and $I(x, y, t-1)$ denote the key and offset frames respectively. Let $(u(x, y), v(x, y))$ denote the true optical flow of pixel (x, y) in the key image. The optical flow can be decomposed as

$$\begin{aligned} u(x,y) &= u_{Z_{ref}}(x,y) + u_p(x,y) \\ v(x,y) &= v_{Z_{ref}}(x,y) + v_p(x,y) \end{aligned} \tag{8}$$

where $(u_{Z_{ref}}, v_{Z_{ref}})$ denotes the flow due to the reference surface Z_{ref} at level l and (u_p, v_p) denotes the parallax due to Z_{ref}. Assuming brightness constancy, we have

$$I(x,y,t) = I(x - u_{Z_{ref}} - u_p, y - v_{Z_{ref}} - v_p, t-1) \tag{9}$$

Assuming a small parallax field, we make the approximation

$$I(x+u_p, y+v_p, t) = I(x-u_{Z_{ref}}, y-v_{Z_{ref}}, t-1) \tag{10}$$

Expanding the left hand side of the above equation in Taylor series around (x, y) and neglecting higher order terms, we have,

$$I_x u_p + I_y v_p + \Delta I = 0 \tag{11}$$

where I_x and I_y denote the spatial image gradients and ΔI represents the difference between the key image and the *warped* offset image according to the reference Z. $u_{Z_{ref}}$ and $v_{Z_{ref}}$ are calculated from the reference Z and motion estimates using (**??**) and the offset image is warped towards the key image using bilinear interpolation. Since we know the camera motion and hence the FOE (x_f, y_f), we can write the parallax field as

$$\begin{aligned} u_p(x,y) &= \beta(x,y) du(x,y) \\ v_p(x,y) &= \beta(x,y) dv(x,y) \end{aligned} \tag{12}$$

where $(du(x,y) = \frac{(x-x_f)}{\sqrt[2]{(x-x_f)^2+(y-y_f)^2}}$, $dv(x,y) = \frac{(y-y_f)}{\sqrt[2]{(x-x_f)^2+(y-y_f)^2}})$ denotes the parallax direction and $\beta(x, y)$ denotes the parallax magnitude for pixel (x, y). Equation (11) then becomes

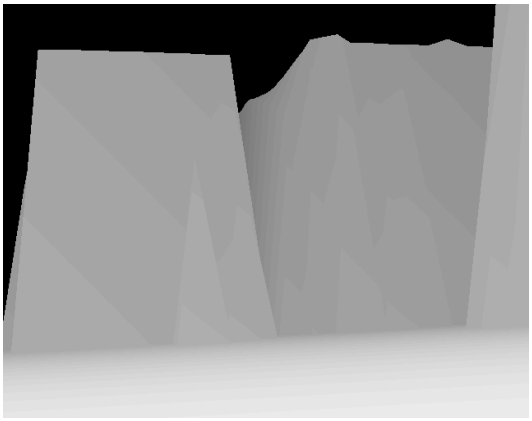
$$\beta(x,y) I_p(x,y) + \Delta I(x,y) = 0 \tag{13}$$

where $I_p = I_x du + I_y dv$ denotes the projection of the intensity gradient in the parallax direction. This is a linear system for each pixel (x, y). Assuming that the parallax magnitude is constant over a neighborhood $N \times N$, for each pixel $(\overline{x},\overline{y})$ we minimize the following error function

$$J(\overline{x},\overline{y}) = \min_p \quad \sum_{(x,y)\in N\times N} < p^T g(x,y) g(x,y)^T p >$$

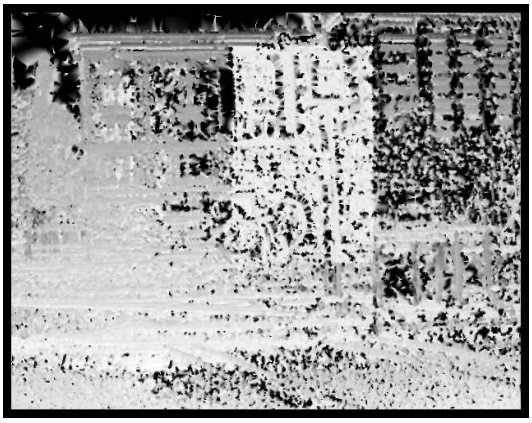
(a)

(b)

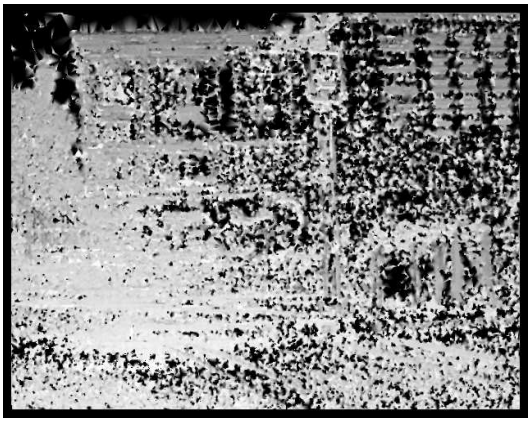


(c)

(d)



(e) L=1



(f) L=3

Figure 3: Real example (a) Key frame (b) Offset frame (c) Reference depth map used for alignment (d) Modified reference depth map (e,f) Estimated depth maps

where $p = \begin{bmatrix} \mu \\ \nu \end{bmatrix}$, $g(x,y) = \begin{bmatrix} I_p(x,y) \\ \Delta I(x,y) \end{bmatrix}$ and $<>$ denotes the smoothing operator defined as

$$< f >= \int_{-\infty}^{\infty} w(x-\overline{x}, y-\overline{y}) f(x,y) dxdy \qquad (14)$$

where w is a smoothing function. Then the parallax magnitude will be given by $\beta(\overline{x},\overline{y}) = \frac{\mu}{\nu}$. To avoid the trivial solution $p = 0$, the constraint $p^T p = 1$ is imposed. Using Lagrange multipliers, the error function can be written as

$$J(\overline{x},\overline{y}) = \min_p \sum_{(x,y)\in N\times N} < p^T g(x,y) g(x,y)^T p > + \lambda(1 - p^T p) \qquad (15)$$

Differentiating with respect to p, we get $Gp = \lambda p$ where

$$G = \begin{bmatrix} < I_p(x,y) I_p(x,y) > & < I_p(x,y) \Delta I(x,y) > \\ < I_p(x,y) \Delta I(x,y) > & < \Delta I(x,y) \Delta I(x,y) > \end{bmatrix}$$

The eigen-vector corresponding to the smaller eigen value of G will be the solution for p from which parallax magnitude can be estimated.

REFERENCES

Agrawal, A. K. and Chellappa, R., 2004: 3D Model Refinement Using Surface-Parallax, *ICASSP*, 285-288.

Bergen, J.R.,Anandan, P., Hanna, K.J. and Hingorani, R., 1992: Hierarchical Model-Based Motion Estimation, *ECCV*,237-252

Horn, B.K.P., 1986: Robot Vision, *McGraw-Hill.*

Kumar, R. and Anandan, P. and Hanna, K.J., 1994: Direct Recovery of Shape from Multiple Views: a parallax based approach, *Proc. 12th IAPR Int'l Conf. Pattern Recognition*, **1**, 685-688.

Trucco, E. and Verri, A., 1998: Introductory Techniques for 3D Computer Vision, *Prentice Hall.*